

# **E.D.N.U.S - Emotion Detection for Individuals with Neurological Disorders Using Speech**



## **EC17713 – MINI PROJECT REPORT**

*Submitted by*

**SURENDARANATH.K**

**(180801201)**

**VISHAL. B**

**(180801223)**

**VISHAL BALAJI SIVARAMAN**

**(180801224)**

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION  
ENGINEERING**

**RAJALAKSHMI ENGINEERING COLLEGE  
CHENNAI – 602105**

**ANNA UNIVERSITY, CHENNAI – 600 025**

**DECEMBER 2021**

# **ANNA UNIVERSITY: CHENNAI 600 025**

## **BONAFIDE CERTIFICATE**

Certified that this project report titled “**Emotion Detection for Individuals with Neurological Disorders Using Speech**” is the bonafide work of “**Surendaranath.K (180801201), Vishal. B (180801223) & Vishal Balaji Sivaraman (180801224)**” who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

### **SIGNATURE**

**Dr. Sheena Christabel Pravin, M.E., Ph.D.,**

### **SUPERVISOR**

Assistant Professor,

Department of Electronics and Communication Engineering,

Rajalakshmi Engineering College,

Thandalam, Chennai – 602 105.

Submitted to Mini Project Viva-Voce Examination held on \_\_\_\_\_

**Internal Examiner**

## ABSTRACT

Human emotion recognition plays an important role in the interpersonal relationship. Emotions are reflected from speech, hand and gestures of the body and through facial expressions. Hence extracting and understanding of emotion is crucial step which has to be undertaken in order to safeguard the interests of the general public as well as the user. Since, emotions are prone to be exhibited from facial expressions, body movement and gestures, and speech. As a result, the technology is said to contribute in the emergence of the so-called emotional or emotive Internet of all the available systems available, we the team propose a novel Speech Emotion Recognition (SER) system where the task is to recognize the emotion from speech irrespective of the semantic contents. Although; emotions are subjective and even for humans it is hard to notate them in natural speech communication, the proposed system outweighs the odds in any given situation. The reason for our choosing would be due to the drawbacks present in the Traditional Facial Recognition system, where the machine makes an erroneous decision with regard to classification of an emotion based on the body language exhibited to the similarity of body language exhibited by another emotion. As disclosed earlier we the team propose a novel hybrid SER system titled **E.D.N.U.S** in order to outweigh the odds. The proposed system **E.D.N.U.S** stands for **E**motion **D**etection for **I**ndividuals with **N**eurological Disorders **U**sing **S**peech

The proposed versatile system incorporates a user interface which gathers the inputs from the user especially the person's audio file, followed by which the concept of feature extraction is performed on the input audio signal features by the proposed autoencoder model, upon which the features are sent to the proposed Super learner model as input for identifying calm, happy, sad and angry emotions, which could be customized based on consumer needs.

## **TABLE OF CONTENTS**

<b>CHAPTER NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
	<b>ABSTRACT</b>	<b>iii</b>
	<b>LIST OF TABLES</b>	<b>vi</b>
	<b>LIST OF FIGURES</b>	<b>vii</b>
<b>1.</b>	<b>INTRODUCTION</b>	<b>1</b>
	1.1 PROJECT OVERVIEW	1
	1.2 TECHNOLOGY STACK USED	2
	1.3 OBJECTIVES	3
<b>2</b>	<b>REVIEW OF LITERATURE</b>	<b>4</b>
<b>3</b>	<b>EXISTING SYSTEM</b>	<b>6</b>
<b>4</b>	<b>METHODOLOGY</b>	<b>7</b>
	4.1 PROCEDURE	7
	4.2 PROCESS OVERVIEW	7
	4.3 PROCESS NOVELTY	9
<b>5</b>	<b>PROPOSED SYSTEM'S ATTRIBUTES</b>	<b>10</b>
	5.1 WORKING PRINCIPLE	10
	5.2 FEATURES	10
	5.3 TARGET AUDIENCE	11

	5.4	PROPOSED ALGORITHM WORKFLOW CHART	11
	5.5	PROPOSED SUPER LEARNER MODEL ARCHITECTURE	12
	5.6	AUTOENCODER ARCHITECTURE	13
<b>6</b>		<b>RESULTS AND DISCUSSION</b>	<b>18</b>
	6.1	AUTOENCODER RESULTS	18
	6.2	SUPER LEARNER RESULTS	18
	6.3	CONFUSION MATRIX	22
	6.4	PROJECT OUTCOME	23
<b>7</b>		<b>CONCLUSION &amp; FUTURE SCOPE</b>	<b>25</b>
	7.1	CONCLUSION	25
	7.2	FUTURE SCOPE	25
<b>8</b>		<b>REFERENCES</b>	<b>26</b>
<b>9</b>		<b>APPENDIX</b>	<b>28</b>
	9.1	CONFERENCE PAPER	28
	9.2	CONFERENCE CERTIFICATE	36

## **LIST OF TABLES**

<b>TABLE NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
3.1	Existing System	6
15.1	Super Learner Results	19

## LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
4.1	Process Overview	8
5.1	Proposed Algorithm Workflow chart	12
5.2	Proposed Super Learner Model Architecture	13
5.3	Autoencoder General Architecture	14
5.4	Proposed Deep Autoencoder: Complete Auto-Encoder Flowchart	15
5.5	Proposed Deep Autoencoder: Encoder Section Flowchart	16
6.1	Deep Autoencoder Results	18
6.2	Accuracy Score and Cohen's Kappa Metric Analysis	20
6.3	F1-Score Metric Analysis	20
6.4	Jaccard Score Metric Analysis	21
6.5	Hamming Loss Metric Analysis	21
6.6	Confusion Matrix for SLM model with Deep Autoencoder	22
6.7	Confusion Matrix for SLM model without Deep Autoencoder	23
6.8	Output Demo Form	23
6.9	Output of the proposed Speech Emotion Detection System	24

# CHAPTER 1

## INTRODUCTION

Emotion is a psychological condition that is linked to the neurological system. It is what a person feels on the inside as a result of the environment in his immediate surroundings. A person's emotions can be sensed in a variety of ways. Tonal characteristics, face expression and body gesture are some of the predominant ways in which one can conclude a person's emotions. Human information processing includes the computation or categorization of emotion based on speech or facial expression. Categorization of emotion based on speech signal shows more accurate results due to the fact that the facial expression would be the same for certain emotions like guilt, fear, lie, etc. and hence as a result during categorization of the emotion, the model would result in erroneous decisions. So based on the above validated conclusion an ideal, cost effective, applicable solution has been devised which is termed as **E.D.N.U.S: Emotion Detection for Individuals with Neurological Disorders Using Speech**

From the above acronym one can conclude that the proposed system would be categorizing the emotion of any individual based on his speech sample (these include talking, short sentences, etc.), Based on the results appropriate actions could be taken for maintaining the wellbeing of the person.

### 1.1 Project Overview

**Artificial Intelligence** is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

**Machine Learning** is defined as the use and development of computer systems that are able to learn and adapt without following explicit instructions,



by using algorithms and statistical models to analyse and draw inferences from patterns in data.

**Deep Learning** is a type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher-level features from data.

**Speech processing** is the study of speech signals and the processing methods of signals. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signals

**Theme:** Artificial Intelligence /Machine Learning/ Deep Learning/ Speech Processing.

## 1.2 Technology Stack Used

- ❖ **Python IDE:** Since Python would be the preferred language for designing/implementation of the algorithm due to wide variety of libraries and support.
- ❖ **Pytorch:** Vital Library required to process and boost the efficiency of the algorithm (Note GPU support is also available based on the host configuration).
- ❖ **Tensor Flow:** Vital Library required to construct and validate the efficacy of the model (Note GPU support is also available based on the host configuration).
- ❖ **Streamlit:** Vital Library required to host the proposed project as a interactive functional website using the same platform Python.
- ❖ **SMS API services** (optional): For alerting the user with respect to the results of the Psychological Analysis Tool (P.A.T).

### 1.3 Objectives

- 1. Short Term objective:** To construct a fully functional system using software, which  
**(Project)** would be compatible with any host system.
  
- 2. Long Term objective:** To construct a unique, interactive Health Bot which would  
**(Product):** be embedded into a cloud service, followed by which the users could access the same using a dedicated Interactive

## **CHAPTER 2**

### **REVIEW OF LITERATURE**

Turker Tuncer *et al.* (2021), in the paper titled, “Accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques”, used a cryptographic structure called a shuffle box to generate features and iterative neighbourhood component analysis to pick them. The suggested technique is divided into three stages: I Tunable Q wavelet transform (TQWT) for multi-level feature creation, (ii) twine shuffle pattern (twine-shuf-pat) for feature production, and (iii) discriminative features are chosen and categorised using iterative neighbourhood component analysis (INCA).

Mustaqeem *et al.* (2021), in the paper titled, “Att-Net: Enhanced emotion recognition system using lightweight self-attention module”, proposed a simple and lightweight deep learning-based self-attention module (SAM) for the SER system in this work. SAM is provided the transitional features map, which efficiently creates the channel and spatial axes attention map with minimal overheads. the team had made the use of a multi-layer perceptron (MLP) in channel attention to extracting global cues and a special dilated convolutional neural network (CNN) in spatial attention to extract spatial info from input tensor.

Sharmeen M Saleem, *et al.* (2021), in the paper titled, “Multimodal Emotion Recognition using Deep Learning”, examined the use of deep learning to recognize emotional signs in multimodal data and compares their applicability based on current research. Multimodal affective computing systems are compared to unimodal solutions since they have a better classification accuracy. The number of emotions seen, characteristics collected, categorization method, and database consistency all affect accuracy.

Fasih Haider *et al.* (2021), in the paper titled, “Emotion recognition in low-resource” focused on emotion identification from speech data in situations when memory and computing resources are limited. One way to achieve this aim is to reduce the number of characteristics used in inductive inference. In this paper, they compare three state-of-the-art feature selection methods: Infinite Latent Feature Selection (ILFS), ReliefF, and Fisher (generalised Fisher score), to our newly suggested feature selection technique called ‘Active Feature Selection’ (AFS). The evaluation is carried out using two standard acoustic paralinguistic feature sets on three emotion recognition data sets (EmoDB, SAVEE, and EMOVO) (i.e. eGeMAPs and emobase).



Md. Shah Fahad *et al.* (2020), in the paper titled, “DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features” used robust epoch recognition from emotional speech to extract emotion-specific epoch-based characteristics, such as immediate pitch, phase, and excitation strength. The combined feature set outperforms the MFCC characteristics, which have been used as a baseline for SER systems in the literature, by 5.07 percent, and state-of-the-art methods by 7.13 percent. The suggested model outperforms state-of-the-art methods by 2.06% when just MFCC characteristics are used.

## CHAPTER 3

### EXISTING SYSTEM

The existing system in general refer to the current techniques/ products employed in real time or available in the market which aims at tackling the designated problem statement. The existing systems for the proposed deep learning-based speech emotion recognition system, are the manual interrogation or the chatting process which may be time consuming, often leads to wrong conclusions and violates privacy. The other alternative system would be the polygraph machine, which is prone to tampering of data. A summary of the existing systems and their drawbacks are presented in Table 3.1.

**Table 3.1 Existing System**

Topic	Picture	Drawbacks
Manual Interrogation /Chatting Process		1) Time Consuming 2) Often leads to wrong Conclusion 3) Violates privacy
Polygraph machine		1) Subjects can bypass the system 2) Tampering of Data is possible

## **CHAPTER 4**

### **METHODOLOGY**

#### **4.1 Procedure**

The proposed deep learning based speech recognition system works by the following steps:

**Step 1:** Initially the subject would upload the audio file recording of his/ her speech.

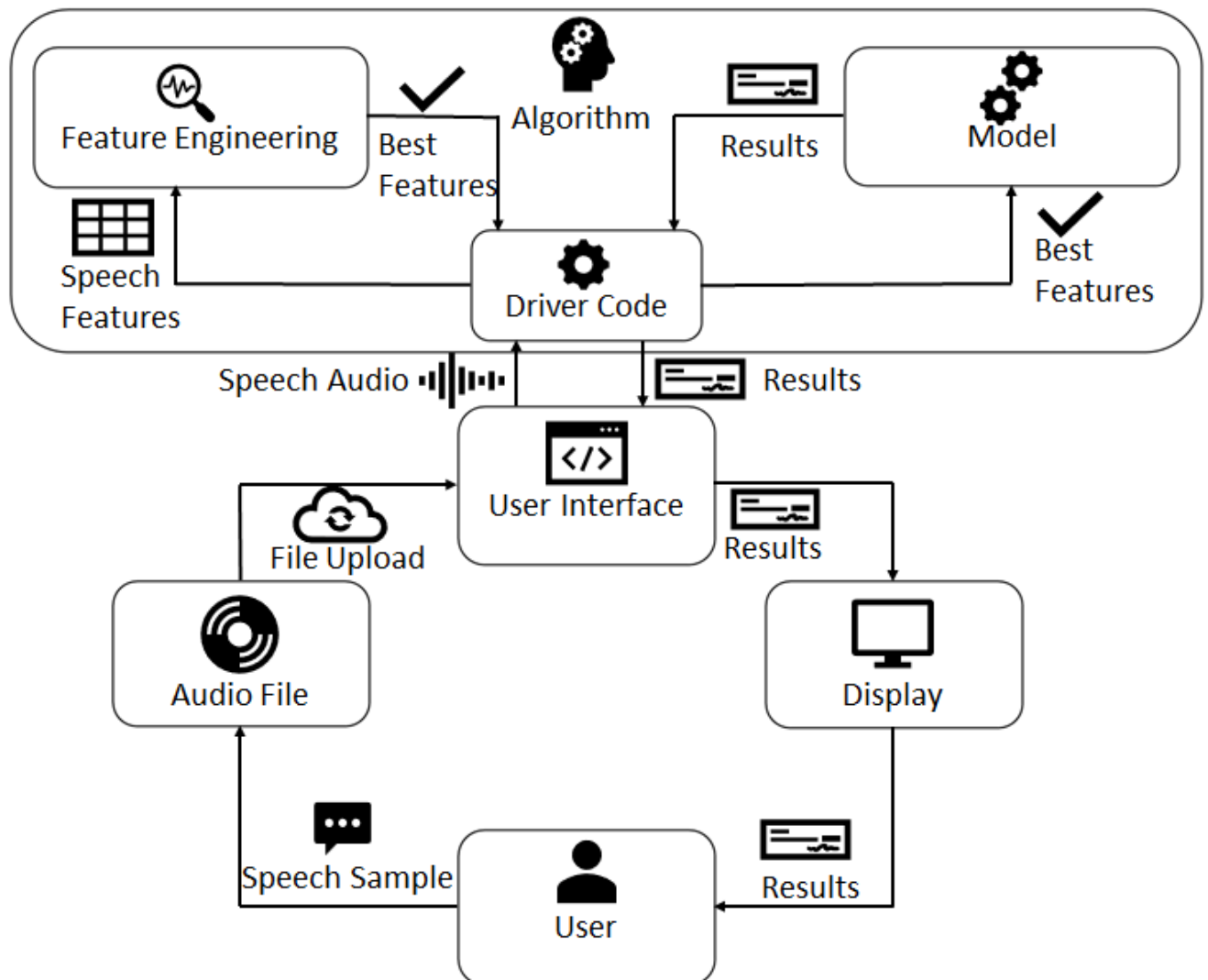
**Step 2:** The speech signal present in the file would be processed into an audio signal that would be sent to the model which is embedded in the User Interface for validation.

**Step 3:** Upon completion of input validation, the proposed autoencoder model would extract the best speech features from the input speech features such as Mel, Chroma & Mel Frequency Cepstral Coefficients (MFCC), followed by which the best features are supplied as input to the proposed Super Learner Model, which would classify the emotion corresponding to the audio signal. Appropriate action may be taken by user for their wellbeing, based on the emotion classification report.

#### **4.2 Process Overview**

As disclosed earlier, the proposed system would require a speech sample from the user in order to perform the thorough review, thus once when the audio file which consists of the speech sample is uploaded through the User Interface (UI) to the system. The system would then execute the process of feature extraction on the input audio file where key speech characteristics encoded in the audio file are detected and retrieved by the speech feature extraction code which is incorporated in the driver code. These speech features are then supplied as input to the proposed Deep Autoencoder which leverages the function of Feature

engineering i.e. the best features are discovered and extracted from the given set of input speech features. These best characteristics are subsequently offered as input to the proposed super learner model for emotion analysis. Once the process of emotion analysis is performed by the super learner model i.e., when the appropriate emotion for the specified input audio file is effectively concluded by the model. The ensuing feeling is eventually conveyed back to the user, upon which measures might be done for the person's wellness. Furthermore, Figure 4.1 renders the whole functional flow of the proposed system



**Fig 4.1 Process Overview**

### **4.3 Process Novelty**

The unique features of the proposed model are as follows:

**1) Multi Label Classification:**

The proposed classification model can classify up to 4 unique/similar emotions with high accuracy.

**2) Feature Engineering:**

Speech Features extracted by a customized neural network framework, termed as the autoencoder is incorporated

**3) Super learner Classification Model:**

The proposed learner model comprises of a stack of eight different unique hyper tuned models. The model as a whole facilitates precise classification as a result yield's high accuracy.



## **CHAPTER 5**

### **PROPOSED SYSTEM'S ATTRIBUTES**

#### **5.1 Working Principle**

As discussed earlier, two pivotal components namely the autoencoder and the super learner model contribute as a backbone in the proposed project.

##### **5.1.1 Autoencoder:**

An Auto encoder is a customizable feed forward neural network which comprises of an encoder and decoder model. Now the primary function of the auto encoder model is to reduce complexity by performing dimensionality reduction (selecting the best features), which in turn would contribute to a boost in accuracy during training of the parent model. In this project a deep Auto encoder have been proposed.

##### **5.1.2 Model:**

The parent model chosen for training would be a super learner model which is more or less a cascaded structure of a number of machine learning models, so since this project focuses on Emotion classification based on speech signal hence, a Super Learner Classification Algorithm is proposed.

#### **5.2 Features**

##### **1) Accessibility:**

The proposed system incorporates a unique interactive website, which could be easily be accessed by users across the globe with the aid of internet.

## **2) Security:**

The proposed system restricts access to the designated user alone, making the data provided by the user feel safe.

## **3) Precision:**

The proposed system as a whole yield better results with high efficiency when compared to results yielded by other models.

## **4) Compatibility:**

The proposed system is highly compatible and can be accessed across a wide range of devices.

## **5) Low Time / Memory Consumption:**

The proposed system would consume low time and space while processing results with high efficiency.

## **6) Customizable:**

The proposed system can be customized based on consumer requirements

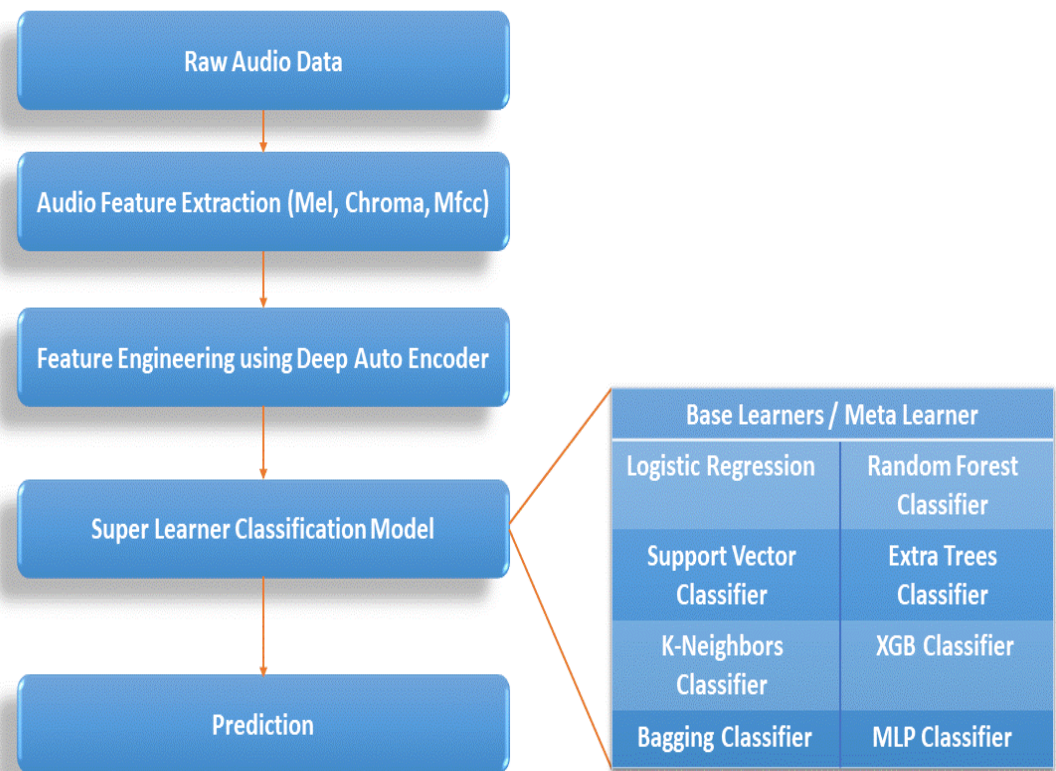
### **5.3 Target Audience**

- 1) Police/Military Police (MP) Interrogations
- 2) Psychiatrist Therapy Sessions
- 3) Company Recruitment (Interview) process
- 4) Employee Stress Test
- 5) Psychological Evaluation Tool (PET) for individuals

### **5.4. Proposed Algorithm Workflow chart**

Fig5.1 depicts the proposed algorithm flowchart incorporated in this project. Initially the audio features Mel, Chroma and MFCC, are extracted from the audio data provided. The features are then fed to the Autoencoder to select

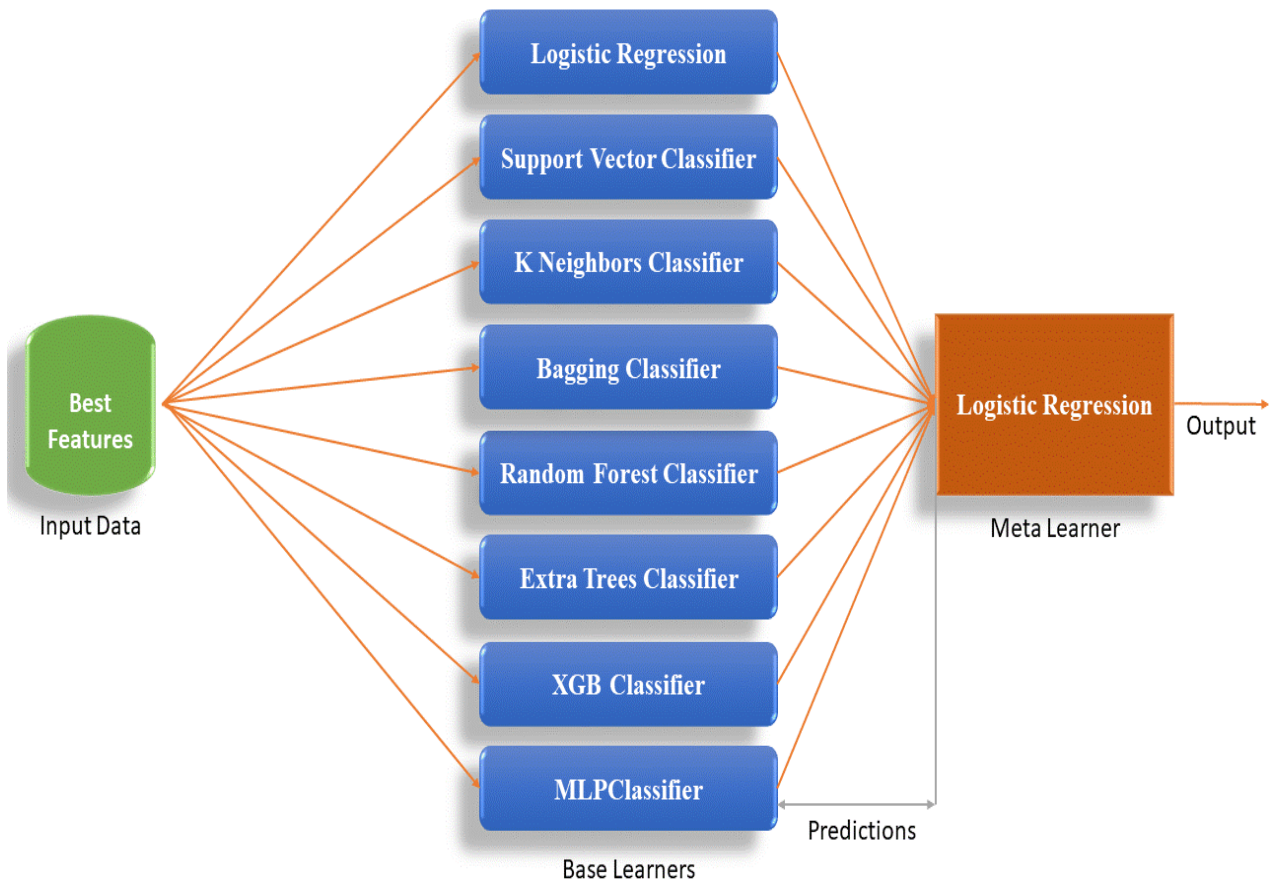
the best features using dimensionality reduction. These selected features are fed to the Super learner model which uses a combination of machine learning and Deep Learning algorithm to finally predict the emotion felt by the provider.



**Fig 5.1 Proposed Algorithm Workflow chart**

## 5.5 Proposed Super Learner Model Architecture

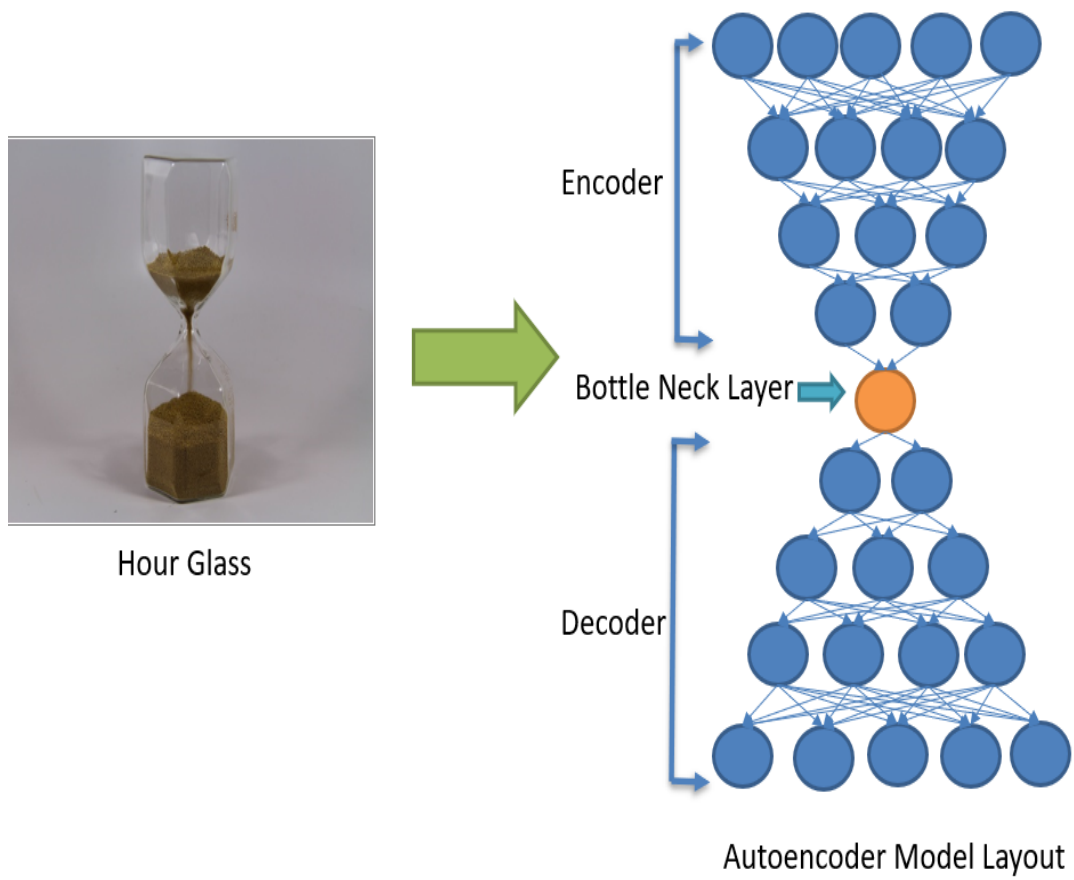
Super Learner is an algorithm that uses cross-validation to estimate the performance of multiple machine learning models, or the same model with different settings. It then creates an optimal weighted average of those models, aka an "ensemble", using the test data performance. The below diagram explains the architecture of the super learner model equipped. The proposed model has 8 base learners and 1 meta learner as depicted in Figure 5.2.



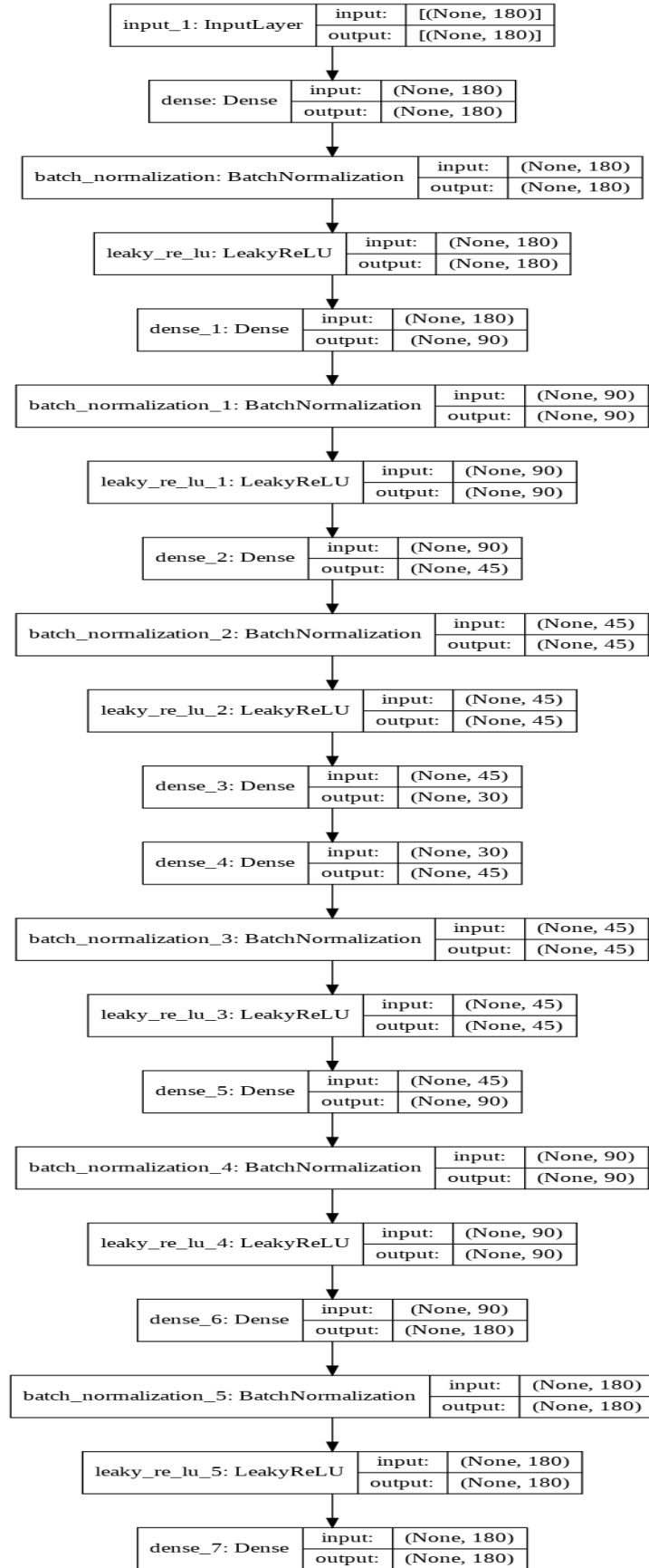
**Fig 5.2 Proposed Super Learner Model Architecture**

## 5.6 Autoencoder Architecture

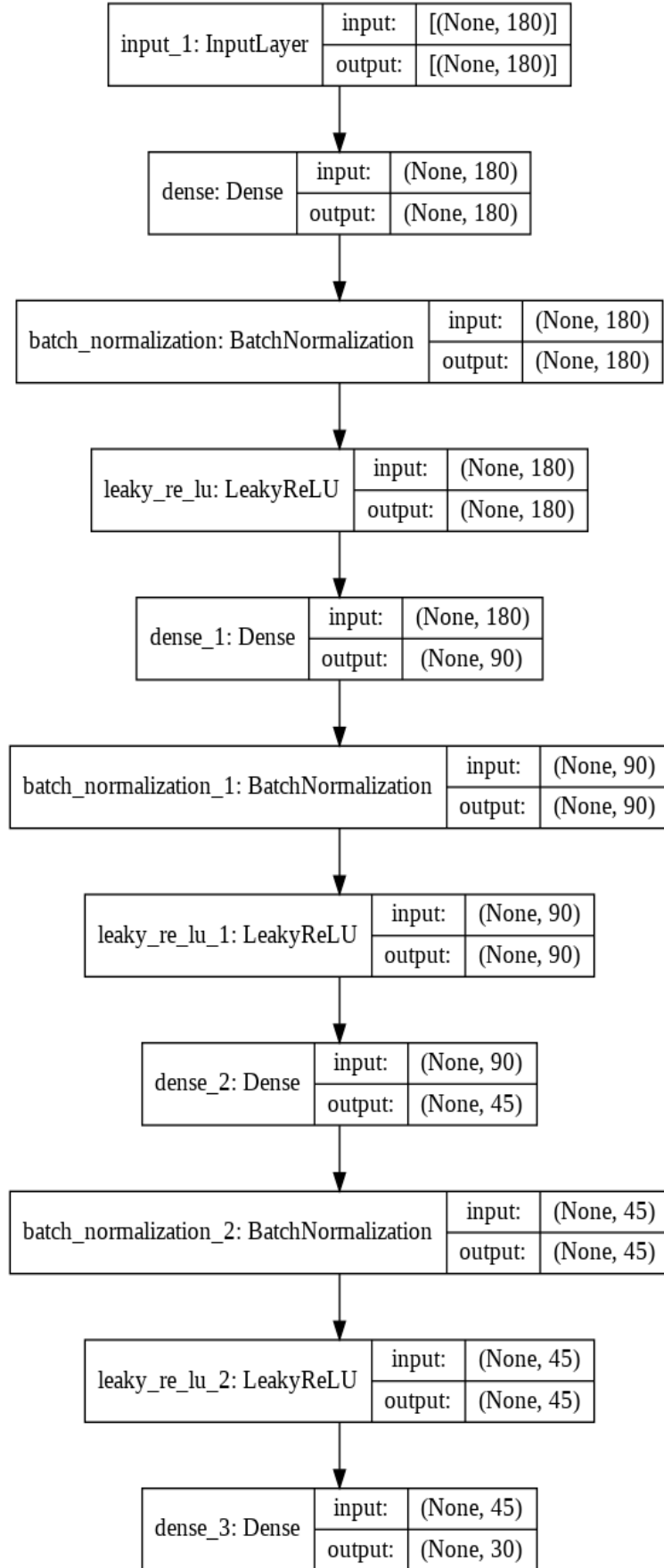
An autoencoder is a neural network architecture capable of discovering structure within data in order to develop a compressed representation of the input. The autoencoder in general comprises of three sections namely encoder, bottle neck and the decoder. The encoder section focuses on shrinking the number of input features fed to the model. The bottle neck section provides the best set of features engineered from the model. Finally, the decoder section aims at reconstructing the features. Furthermore, in general the structure of an autoencoder bears close resemblance with that of an hour glass as depicted in Figure 5.3.



**Fig 5.3 Autoencoder General Architecture**



**Fig 5.4 Proposed Deep Autoencoder: Complete Auto-Encoder Flowchart**



**Fig 5.5 Proposed Deep Autoencoder: Encoder Section Flowchart**

Furthermore, the Figure 5.4 renders the complete flowchart of a seven-layer Deep autoencoder network where the first three layers are dedicated to serve the functions of an encoder. Similarly, the last three layers are dedicated to serve the functions of a decoder. The middle layer which consists of 30 neurons is dedicated to serve as the Bottle Neck layer. In general, the neurons in the first layer of the encoder section are adjusted to the number of features (180). Similarly, the neurons in the final layer of the decoding portion are adjusted to the number of features (180). In addition, all the layers' activation functions have been modified to Rectified Linear Unit (ReLU) with an addition of special functions such as Batch Normalization and Leaky ReLu respectively. Figure 5.5 represents the flowchart of the proposed autoencoder's encoder section.



## CHAPTER 6

### RESULTS AND DISCUSSION

#### 6.1 Autoencoder Results

As depicted in Figures 6.1, the proposed Deep Autoencoder model achieves a training score of 0.9948 and a validation score of 0.9558, with minimal training and validation loss of 0.05 for a total of 10,000 epochs respectively.

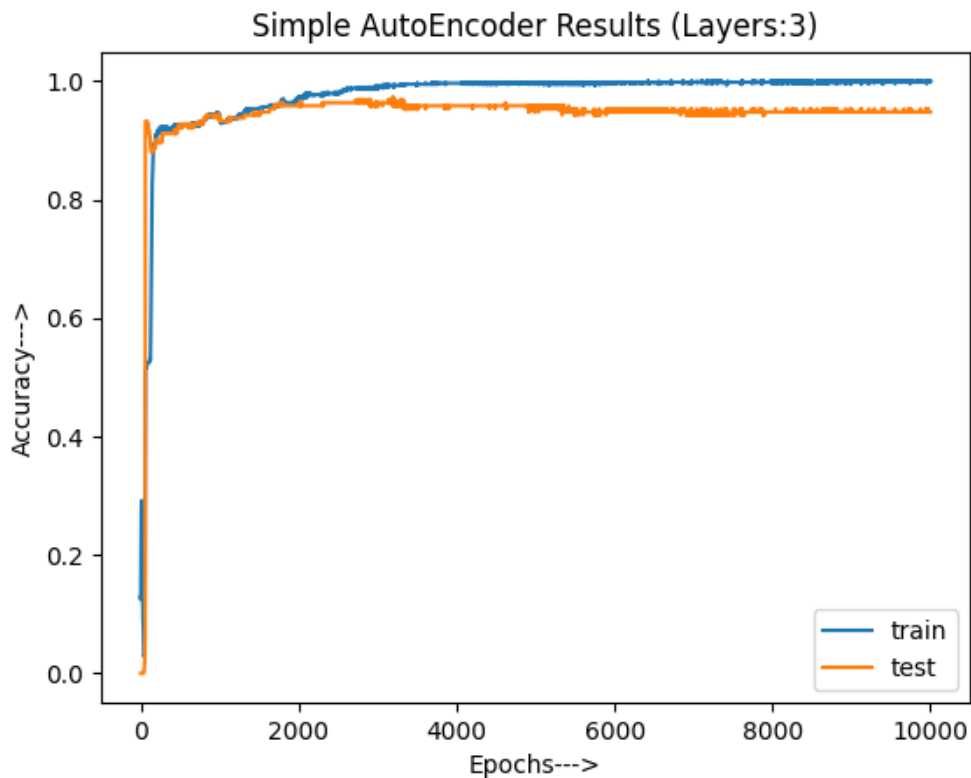


Fig 6.1 Deep Autoencoder Results

#### 6.2 Super Learner Results

The proposed model results were documented and analyzed in a tabular and graphical format respectively.

##### 6.2.1 Super Learner Tabulation Analysis

The table 15.1 depicts the Super Learner Results of the model with and without using deep autoencoder.

**Table 15.1 Super Learner Results**

<b>Metric</b>	<b>SLM with Deep Autoencoder</b>	<b>SLM without Deep Autoencoder</b>
<b>Accuracy Score</b>	84	76
<b>Balanced Accuracy Score</b>	83	76
<b>Cohen Kappa Score</b>	78	68
<b>F1 Score (Macro)</b>	83	76
<b>F1 Score (Micro)</b>	84	76
<b>F1 Score (Weighted)</b>	84	76
<b>Jaccard Score (Macro)</b>	72	61
<b>Jaccard Score (Micro)</b>	72	61
<b>Jaccard Score (Weighted)</b>	72	62
<b>Hamming Loss</b>	0.1614	0.2395

### 6.2.2 Super Learner Graphical Analysis

The images rendered below depict the various Super Learner metrics of the model with and without using Deep Autoencoder. Figure 6.2 depicts the Accuracy and Cohen's Kappa score analysis of the proposed models. Furthermore, Figure 6.3 & 6.4 depicts the F1-Score metrics and Jaccard Score metric analysis of the proposed models while Figure 6.5 depicts the hamming loss analysis rendered by the models respectively.

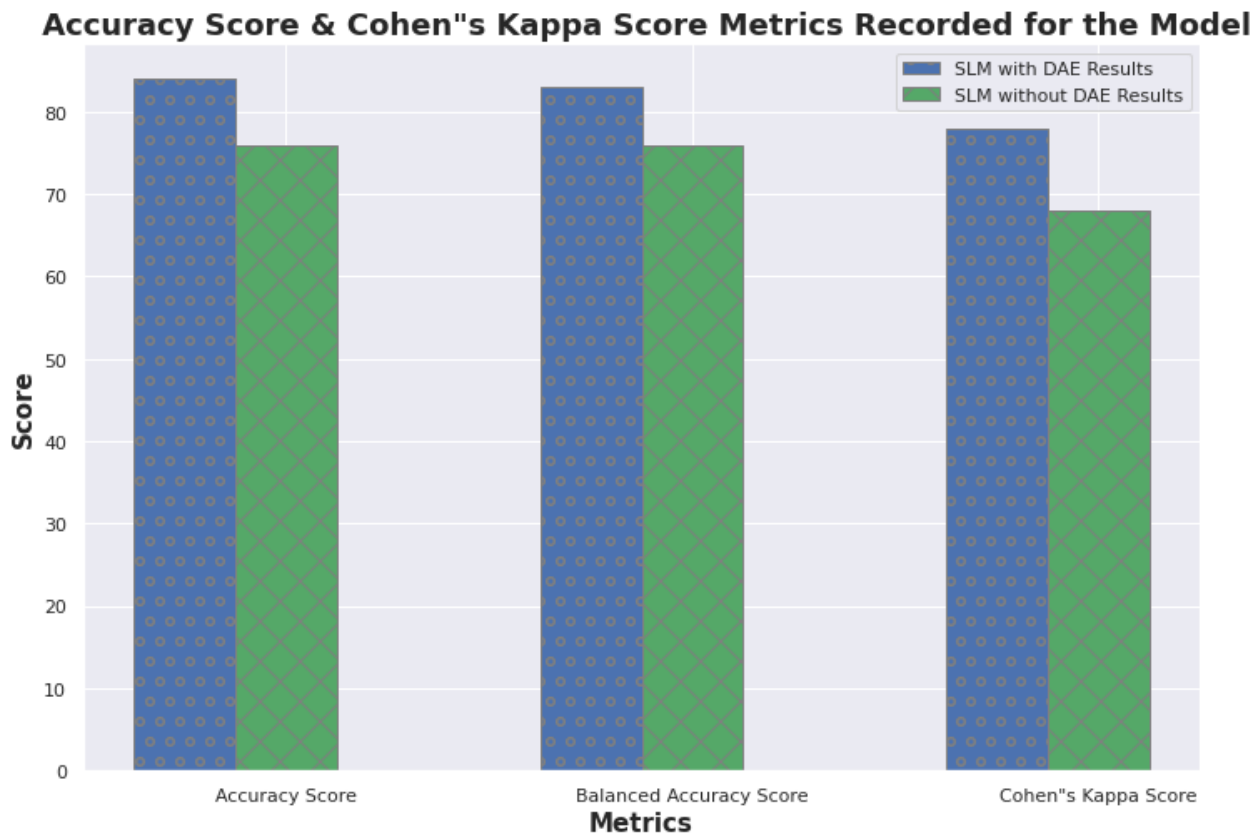


Fig 6.2 Accuracy Score and Cohen's Kappa Metric Analysis



Fig 6.3 F1-Score Metric Analysis

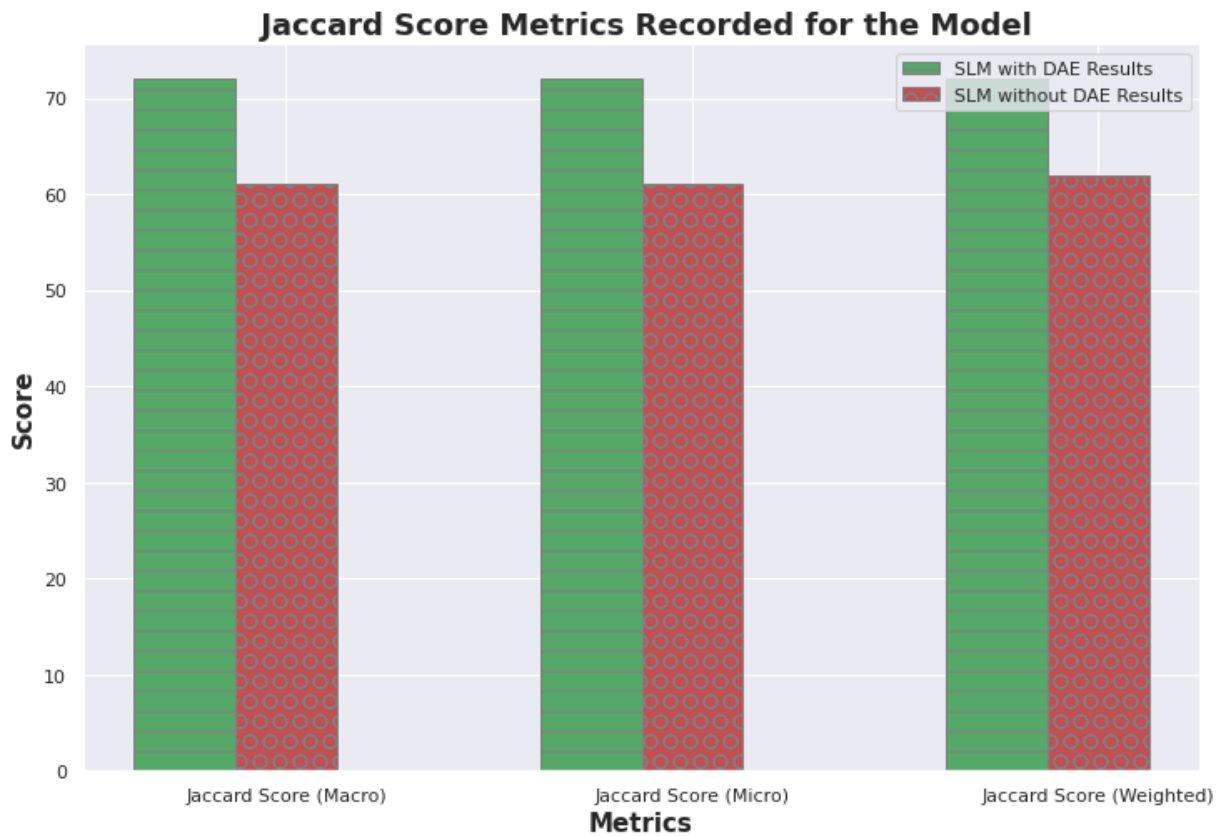


Fig 6.4 Jaccard Score Metric Analysis

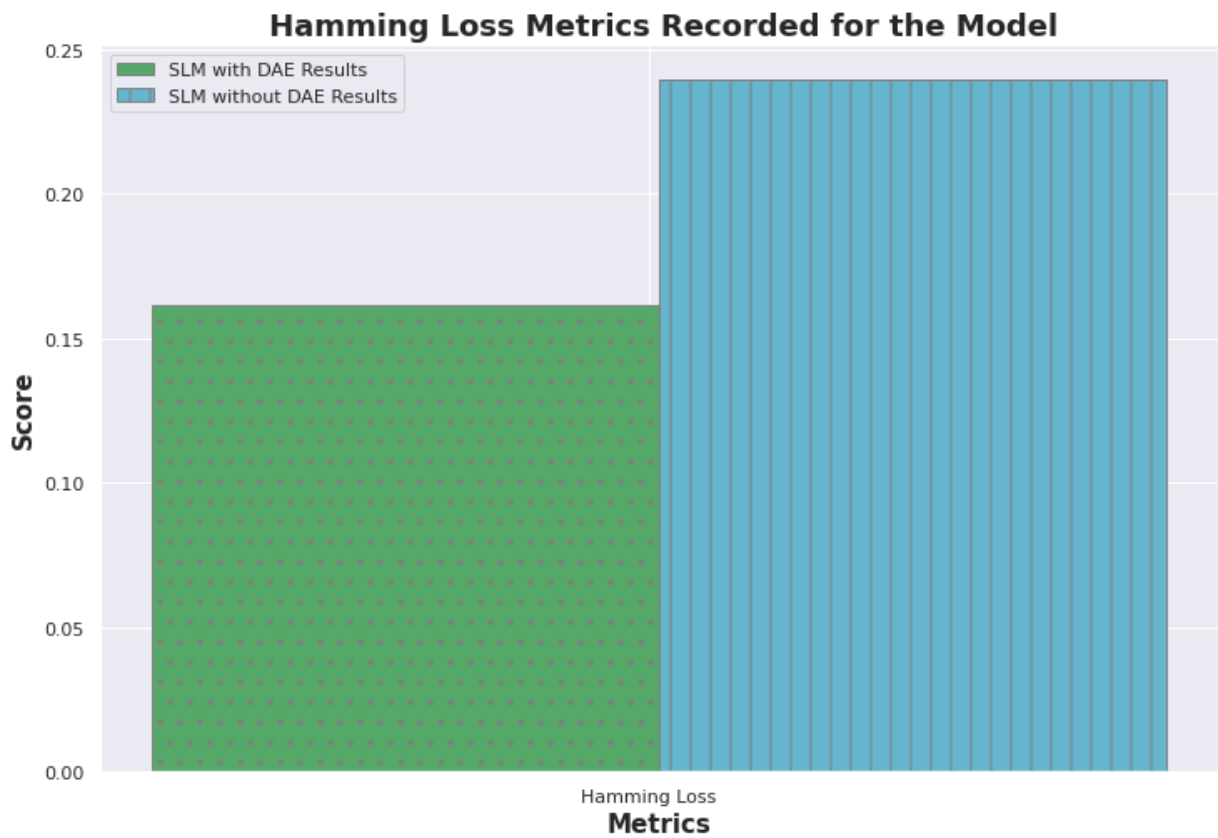
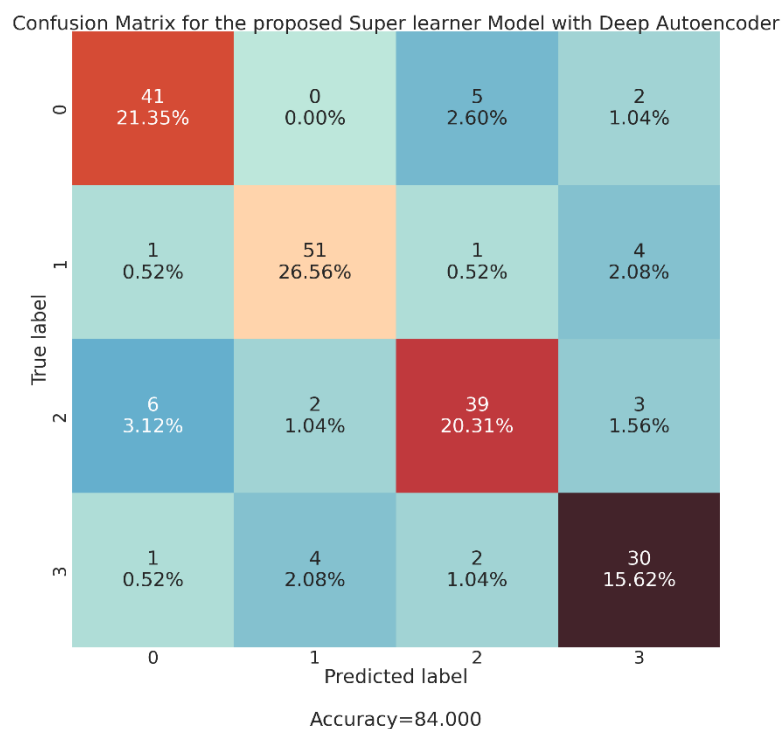


Fig 6.5 Hamming Loss Metric Analysis

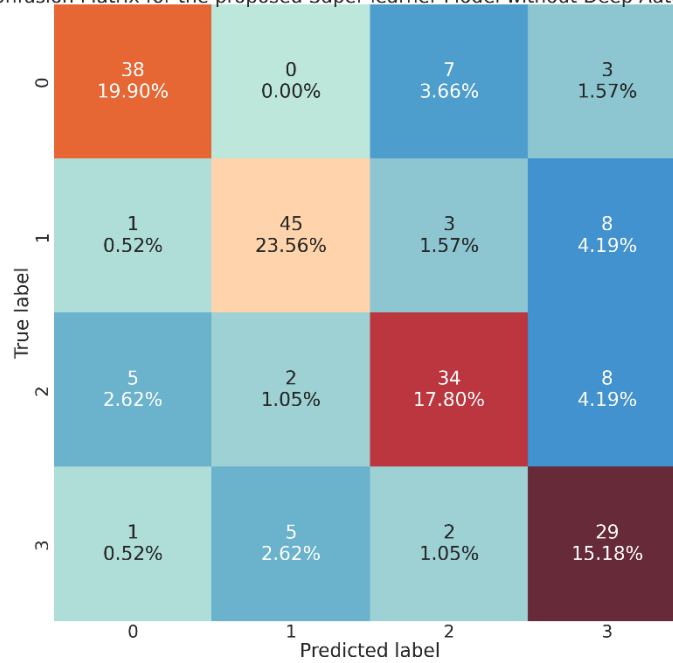
### 6.3 Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. The Figures 6.2 and Figure 6.3 shows the Confusion matrix of SLM Model with and without Deep Autoencoder respectively.



**Fig 6.6 Confusion Matrix for SLM model with Deep Autoencoder**

Confusion Matrix for the proposed Super learner Model without Deep Autoencoder



Accuracy=76.000

**Fig 6.7 Confusion Matrix for SLM model without Deep Autoencoder**

## 6.4 Project Outcome

Choose your desired option from the available maintopics

Product Demo

Choose your desired choice from the available subtopics under Technical Information

Proposed Flow

Choose your desired choice from the available subtopics under Project Information

Introduction

### Product Demo Form

Kindly enter the details with care

Enter your full name

Vishal

6/30

Enter your age

18

- +

Enter your phonenumber

9500078205

10/10

Enter your email id

vbs.official@gmail.com

21/50

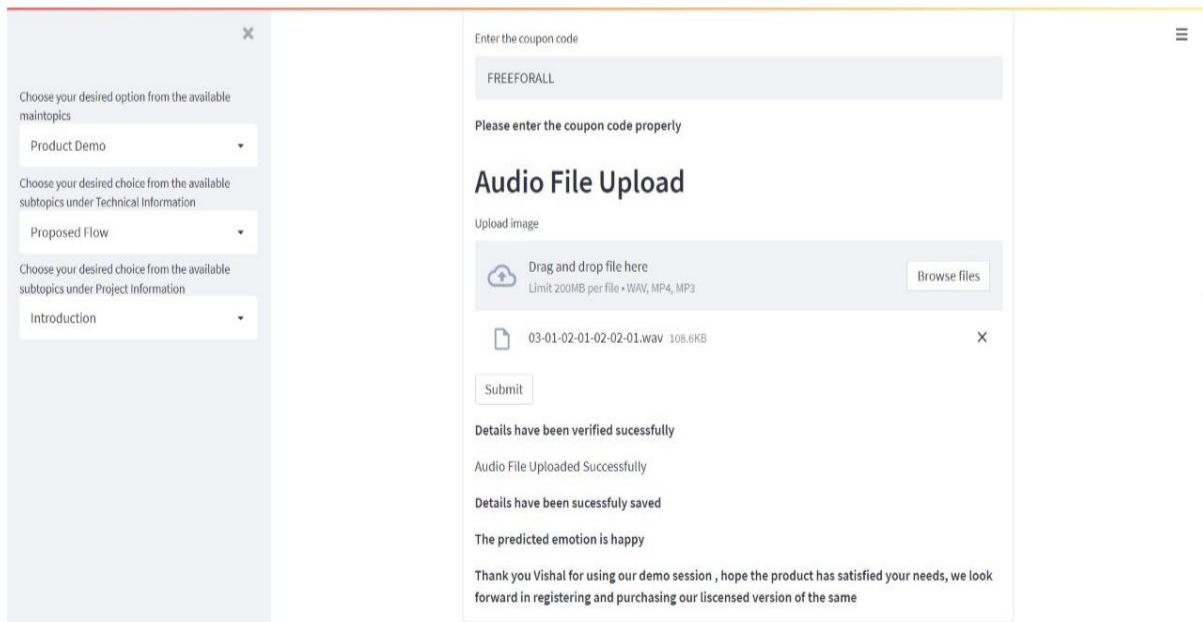
klm234der

Enter the coupon code

klm234der

Please enter the coupon code properly

**Fig 6.8 Output Demo Form**



**Fig 6.9 Output of the proposed Speech Emotion Detection System**

The Figures 6.8 & 6.9 portray the comprehensive functioning of the suggested system. Concurrently, the figures also demonstrate the User interface (UI) characteristics incorporated in the system in order to promote improved user experience.

## CHAPTER 7

### CONCLUSION & FUTURE SCOPE

#### 7.1 Conclusion

Therefore, it can be concluded that the proposed Super learner model incorporated with a Deep Autoencoder model outperforms other models with its ground breaking results on a dense dataset like speech. As a result, based on the above set of conclusive results it can be positively concluded that the proposed project has been implemented based on the features rendered earlier Furthermore, It is proposed to take this project forward as a product by incorporating certain features elaborated under the future scope section.

#### 7.2 Future Scope

A fully functional system using Raspberry pi which would be compatible with any host system may be constructed. The system may be brought out as a product namely, **E.D.N.U.S: Emotion Detection for Individuals with Neurological Disorders Using Speech**



## **CHAPTER 8**

### **REFERENCES**

- [1] Stuti Juyal, Chirag Killa, Gurvinder Pal Singh, Nishant Gupta, Vedika Gupta ‘Emotion Recognition from Speech Using Deep Neural Network’  
[https://link.springer.com/chapter/10.1007/978-3-030-76167-7\\_1](https://link.springer.com/chapter/10.1007/978-3-030-76167-7_1)
- [2] Huang, F., Zhang, J., Zhou, C. et al. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. *Landslides* 17, 217–229 (2020). DOI: 10.1007/s10346-019-01274-9  
<https://link.springer.com/article/10.1007%2Fs10346-019-01274-9>
- [3] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in *IEEE Access*, vol. 7, pp. 117327-117345, 2019, DOI: 10.1109/ACCESS.2019.2936124 <https://ieeexplore.ieee.org/document/8805181>
- [4] Meheebub Sahana, Binh Thai Pham, Manas Shukla, Romulus Costache, Do Xuan Thu, Rabin Chakraborty, Neelima Satyam, Huu Duy Nguyen, Tran Van Phong, Hiep Van Le, Subodh Chandra Pal, G. Areendran, Kashif Imdad & Indra Prakash (2020) Rainfall induced landslide susceptibility mapping using novel hybrid soft computing methods based on multi-layer perceptron neural network classifier, *Geocarto International*, DOI: 10.1080/10106049.2020.1837262
- [5] Sheena Christabel Pravin, Palanivelan, M, ‘A Hybrid Deep Ensemble for Speech Disfluency Classification’, *Circuits, Systems, and Signal Processing*, Springer, vol. 40, no.8, pp. 3968-3995, July 2021.  
[https://www.researchgate.net/publication/349228170\\_A\\_Hybrid\\_Deep\\_Ensemble\\_for\\_Speech\\_Disfluency\\_Classification](https://www.researchgate.net/publication/349228170_A_Hybrid_Deep_Ensemble_for_Speech_Disfluency_Classification)

- [6] Krishnan, P.T., Joseph Raj, A.N. & Rajangam, V. Emotion classification from speech signal based on empirical mode decomposition and non-linear features. *Complex Intell. Syst.* 7, 1919–1934(2021).DOI:10.1007/s40747-021-00295-z <https://link.springer.com/article/10.1007%2Fs40747-021-00295-z>
- [7] Mohamad Nezami, O., Jamshid Lou, P. & Karami, M. ShEMO: a large-scale validated database for Persian speech emotion detection. *Lang Resources & Evaluation* 53, 1–16 (2019).DOI:10.1007/s10579-018-9427-x <https://link.springer.com/article/10.1007%2Fs10579-018-9427-x>
- [8] M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 858-862, DOI: 10.1109/ISS1.2017.8389299 <https://ieeexplore.ieee.org/document/8389299>
- [9] M. N. Stolar, M. Lech, R. S. Bolia and M. Skinner, "Real-time speech emotion recognition using RGB image classification and transfer learning," 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS), 2017, pp. 1-8, DOI: 10.1109/ICSPCS.2017.8270472 . <https://ieeexplore.IEEE.org/document/8270472>
- [10] J. D. Arias-Londono, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, On combining information from modulation spectra and Mel-frequency cepstral coefficients for automatic detection of pathological voices, *Logoped. Phoniatr. Vocol.* 36(2) (2011) 60–69. <https://pubmed.ncbi.nlm.nih.gov/21073260/>

# APPENDIX

## Sparse Autoencoder based Speech Emotion Recognition

Vishal Balaji Sivaraman<sup>1</sup>, Sheena Christabel Pravin<sup>1\*</sup>, Surendaranath.K<sup>1</sup>, Vishal.A<sup>1</sup>, M. Palanivelan<sup>1</sup>, Saranya.J<sup>1</sup>, Priya L<sup>2</sup>

<sup>1</sup>Department of ECE, Rajalakshmi Engineering College, Chennai, India

<sup>2</sup>Department of Information Technology, Rajalakshmi Engineering College, Chennai, India

vishalbalaji.sivaraman.2018.ece@rajalakshmi.edu.in

sheena.s@rajalakshmi.edu.in, ORCID no. 0000-0001-8520-3322.

surendaranath.k.2018.ece@rajalakshmi.edu.in

[vishal.a.2018.ece@rajalakshmi.edu.in](mailto:vishal.a.2018.ece@rajalakshmi.edu.in)

[palanivelan.m@rajalakshmi.edu.in](mailto:palanivelan.m@rajalakshmi.edu.in), ORCID no. 0000-0001-8278-9348

[saranya.j@rajalakshmi.edu.in](mailto:saranya.j@rajalakshmi.edu.in), ORCID no. 0000-0001-5991-389x

[priya.l@rajalakshmi.edu.in](mailto:priya.l@rajalakshmi.edu.in), ORCID no. 0000-0003-4995-1993

**Abstract.** One of the most natural methods for humans to express themselves is through speech. People nowadays are drawn to alternative ways of communication including emails, text messages, and the usage of emoticons to express their feelings. Given the importance of emotions in communication, recognizing and analyzing them is crucial in today's digital age of remote communication. As emotions are complicated, recognizing them can be difficult. There is no universally accepted method for quantifying or categorizing them. A Speech Emotion Recognition system is defined as a set of methods for processing and categorizing speech signals to detect the emotions inherent within them. This work proposed a hybrid model namely the Sparse AutoEncoder-Multi-Layered Perceptron (SAE-MLP) model. The SAE model is used for feature extraction and the MLP for the categorization of speech emotions.

**Keywords:** Speech, Emotion, Features, Classification, Sparse AutoEncoder, Multi-Layered Perceptron.

## 1 Introduction

There are three types of characteristics in speech: lexical characteristics such as the vocabulary, visual characteristics viz. the speaker's expressions, and auditory characteristics like pitch, tone, and jitter. One or more of these attributes can be used to address the challenge of speech Emotion Recognition (SER). If emotions were to be predicted from real-time audio, a transcript of the speech would be required, which would demand a second stage of text extraction from speech. Further visual analysis requires access to video recordings of conversations that are not available in all situations, but there is a significant demand for audio, allowing acoustic analysis to be performed in real time while the conversation is ongoing because there is a significant need for audio data to complete the task. As a result, it was decided to concentrate on the acoustic components of speech. Furthermore, there are also two methods of representing emotions:

- Discrete Classification: Sorting emotions under discrete labels such as angry, happy, surprise, fear, etc.
- Dimensional Representation: Emotions are represented using variables such as valence, with negative and positive values, activation or vitality function, which expands from low to high values, along with dominance, which ranges from active to passive scale.

Both systems have pluses and minuses. The representation using dimension, is complex and provides details for prediction, but it is also harder to execute and shortage of annotated audio data. The discrete approach is more straightforward and simpler to execute, but it lacks the predictive context that the dimensional representation provides. The discrete classification strategy was used in this investigation due to a lack of dimensionally annotated data in the public domain.

The use of speech to recognize emotions has gained popularity in the scientific community. Several efforts have been done earlier by numerous scholars throughout the world on the same subject. Identification of eight

different emotions [1] has been experimented by extracting various speech features viz. Fundamental frequency, Energy, Mel-Frequency Cepstral Coefficients, Mel Energy Spectrum Dynamic Coefficients with an accuracy of 50%. Convolutional Neural Networks (CNN) based speech emotion detection system [1] that could recognize 6 classes of emotion. A speech emotion recognition system [2] that recognizes emotions from one-second frame of raw speech spectrograms, by training a deep learning network was built with the eNTERFACE database and the Surrey Audio-Visual Expressed Emotion database as training datasets. In the recent literature, hybrid machine learning models [3] have found popularity owing to their greater performance and reduced complexity. This research work also aims to introduce one such hybrid model for speech emotion recognition.

### 1.1 Contributions

The significant contributions of this research paper are as follows:

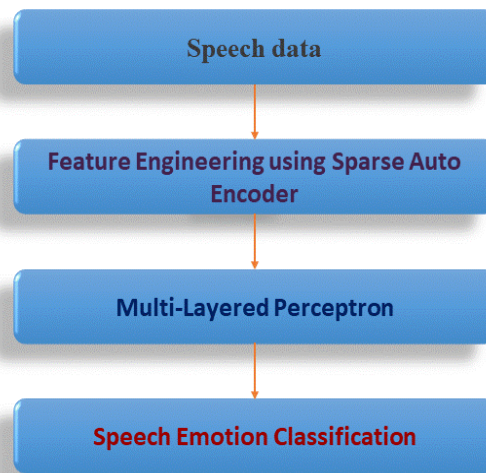
- A hybrid model namely the SAE-MLP model is proposed for speech emotion recognition
- Model regularization has been introduced in the form of drop out at each layer of the Sparse Autoencoder to avoid over-fitting
- Fine hyper-parameter tuning has been introduced to make the proposed model efficient in categorizing speech emotions with high accuracy and precision.

## 2 Speech Emotion Dataset

In this research, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was employed. The collection contains 7356 audio files. Twenty four professional actors in the database vocalize two lexically matched sentences in a neutral American accent. Speech has varied expressions, including the calm, happy, sad, angry, terrified, surprise and disgust emotions, whereas music contains peaceful, happy, sad, angry, and scary emotions. There are two levels of emotional intensity such as normal and strong for each expression and a neutral expression.

## 3 Proposed SAE-MLP Model

The proposed model workflow is portrayed in the schematic shown in Figure. 1. Initially, the raw speech input signal should be provided to the Sparse autoencoder which performs automatic feature engineering. The encoder part of the SAE creates a new latent depiction of the input speech signal while the decoder reconstructs the latent features to bring out the same speech signal. After adequate training, the decoder of the SAE is removed and the latent representations, which have reduced dimensionality, are provided as input to the Multi-Layer Perceptron (MLP) classifier. The overall flow diagram of the proposed framework is given in Figure 1.



**Figure 1.** Pipeline structure for final prediction

The MLP [3] classifies the signals into definite speech emotion classes after being trained on the latest features from the SAE. Thus, the Sparse AutoEncoder acts as a generative model while the MLP functions as the discriminative model. A detailed description of the proposed SAE-MLP model is pictured in Figure 2.

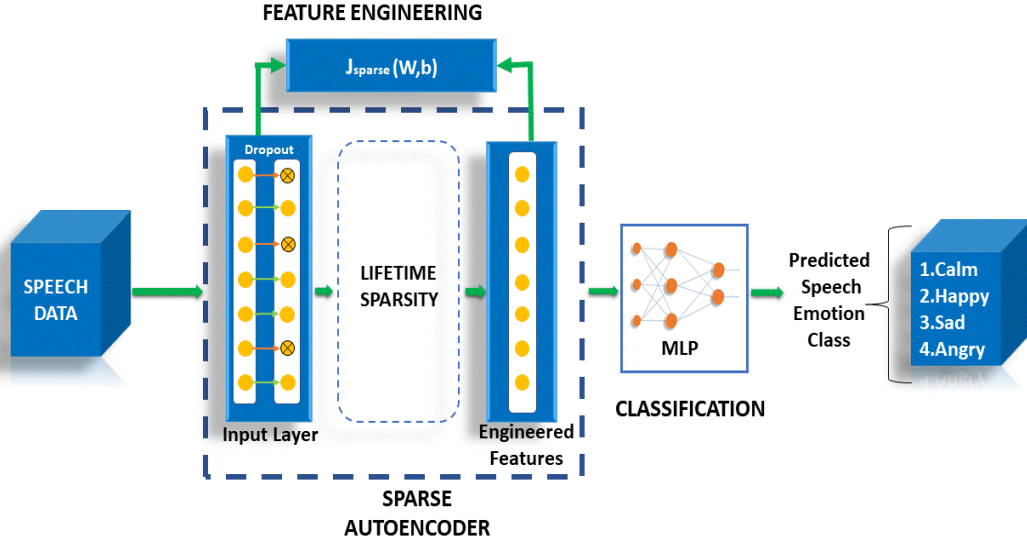


Figure 2. Proposed SAE-MLP model architecture

### 3.1 Train/Test Data Allocation

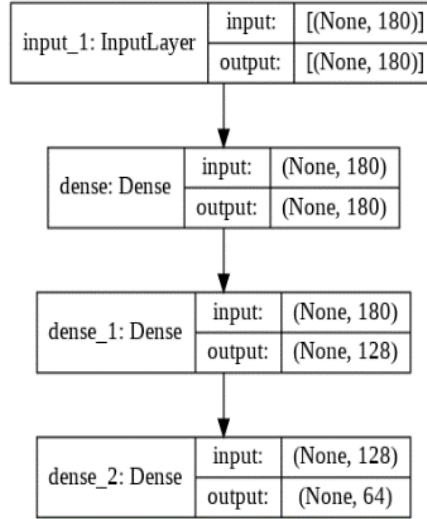
The complete dataset is divided in a ratio of 70:30, with 70% of the dataset being used for model training and 30% for model validation. This method [4] is widely used for training and validating the proposed Sparse Autoencoder and the model.

### 3.2 Feature Engineering Using Sparse Autoencoder

The Sparse Autoencoder model was utilized to reduce dimensionality, resulting in the best features being allocated to the suggested pipeline model for emotion analysis. The proposed autoencoder model was trained over 10,000 iterations during the feature engineering phase.

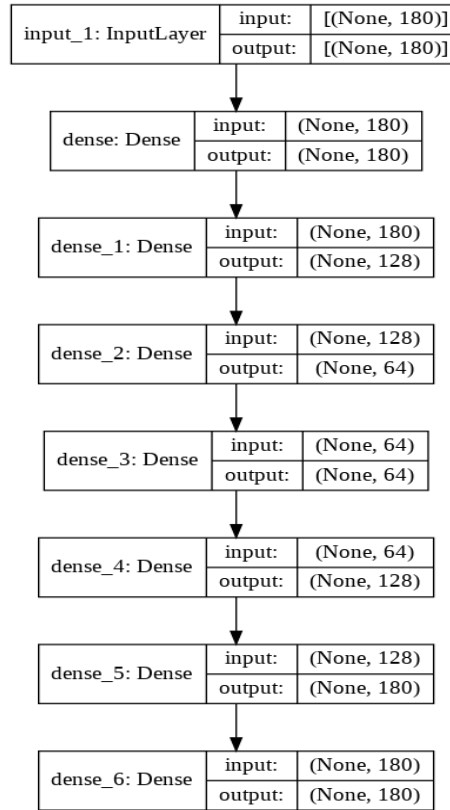
### 3.3 Model Summary of a Sparse AutoEncoder

The proposed Sparse Autoencoder model comprises an encoder and decoder section, both of which are made up of three layers of dense networks each with a distinct count of neurons.



**Figure 3.** Sparse Autoencoder's Encoder block Model Summary

The neurons in the first layer of the encoder section are adjusted to the number of features (180), as shown in Figure 3. Similarly, the neurons in the final layer of the decoding portion are adjusted to the number of features (180). In addition, all layers' activation functions have been modified to Rectified Linear Unit (ReLU) with a constant learning rate, as shown in Figure 4.

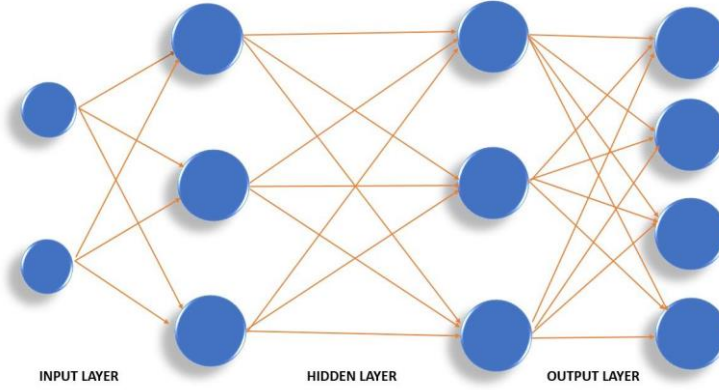


**Figure 4.** Sparse Autoencoder Model Summary

### 3.4 Multi-Layer Perceptron (MLP)

Multi-layer perceptron (MLP) is an extension of artificial neural network (ANN) as shown in Figure 5. The term MLP is ambiguous; it is applicable to any feedforward ANN [5] or networks composed of many layers of perceptron with threshold activation. MPL is three-layered with an input layer, hidden layer, followed by an output

layer. Each node is a neuron with a non-linear activation function, with an exception of the input neurons. Backpropagation is a supervised learning method used by MLP during training. The multi-layered MLP with a non-linear activation function helps to distinguish it from linear perceptrons. MLP excels at discriminating data that is not linearly divisible. The Multi-Layer Perceptron consists of two types of models that follow the same mechanism, one for Regression (MLP Regressor) and the other for Classification (MLP Classifier). As a result, the classification model (MLP Classifier) has been chosen for examination and substantiation of the earlier assertion.



**Figure 5.** Multi-Layer Perceptron Architecture

#### 4 Model Evaluation

SAE model evaluation was performed using accuracy and loss metrics for each epoch, whereas for the MLP Classifier model, score metrics such as Classification Accuracy Score, Cohen Kappa coefficient, Balanced Accuracy Score, F1 Score, and Jaccard Score were used [5], similarly in the case of loss metrics hamming loss metric was used, and finally confusion matrix was drafted to determine the individual class predominance. The metrics are computed using equations (1) to (6).

$$Jaccard\ Score = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

$$Cohen\ Kappa\ Score = \frac{P_0 - P_e}{1 - P_e} \quad (2)$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \quad (4)$$

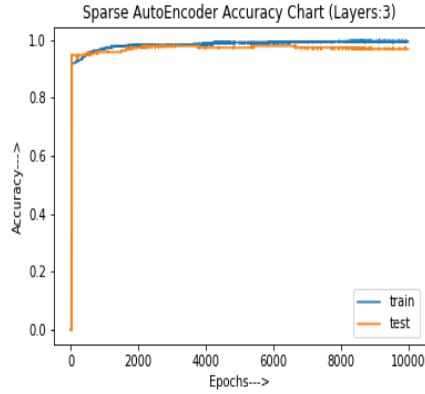
$$Classification\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Hamming\ Loss = \frac{1}{nL} \sum_{i=1}^n \sum_{j=1}^L [I(y_j^{(i)} \neq y'_j^{(i)})] \quad (6)$$

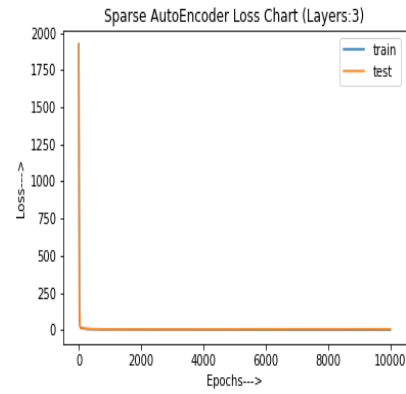
The Jaccard score is calculated by substituting the set of labels predicted by the model and the actual set of labels in equations (1), (2) were used for computing the Cohen Kappa Score for the model upon substitution of  $P_0$  and  $P_e$ , where  $P_0$  is the ratio of observed agreement and  $P_e$  is the expected agreement when both annotators assign labels at random. The F1 score for the model is quantified by substituting the values of Precision and Recall in equations (3) (4) were used for reckoning the Balanced accuracy score for the model, upon substituting the values of Sensitivity and Specificity. Substituting the components True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) into the equation (5), yields the accuracy score for the model. Aside from that, the Hamming loss for a model was enumerated using the set of labels predicted by the model and the actual set of labels, as illustrated in equation (6). Finally, a confusion matrix for the model is drafted by, utilizing the components required in computing a model's accuracy score.

#### 4.1 Sparse Autoencoder Evaluation

As depicted in Figures 6 and 7, the proposed autoencoder model achieves a training score of 0.9948 (99.48 percent) and a validation score of 0.9688 with minimal training and validation loss of 0.05.



**Figure 6.** Sparse Autoencoder Accuracy Chart



**Figure 7.** Sparse Autoencoder Loss Chart

#### 4.2 Multi-Layer Perceptron (MLP) Classifier with and without Sparse Autoencoder results

An experimentation on the efficacy of the MLP model with and without the SAE model was executed. The MLP's performance was well-enhanced when the SAE-based latent features were used to train it. The performance of MLP dipped when the SAE model was removed. The results rendered in Table 1 & 2, represents the results concerning the inclusion and exclusion of sparse autoencoder model with the MLP classifier model. Figures 8 & 9 represent the confusion matrix heatmap for the respective models.

**Table 1.** Model Results

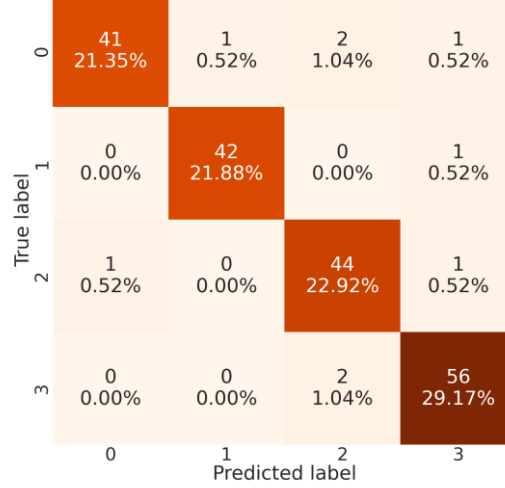
Metrics	MLP Classifier with Sparse Encoder (%)	MLP Classifier without Sparse Encoder (%)
Classification Accuracy Score	95	55
Balanced Accuracy Score	95	54
F1 Score (Macro)	95	51
F1 Score (Micro)	95	55
F1 Score (Weighted)	95	52



**Table 2.** Model Results

Metrics	MLP Classifier with Sparse Encoder (%)	MLP Classifier without Sparse Encoder (%)
Cohen Kappa Score	94	39
Jaccard Score (Macro)	91	36
Jaccard Score (Micro)	91	38
Jaccard Score (Weighted)	91	37
Hamming Loss	4.6875	45.3125

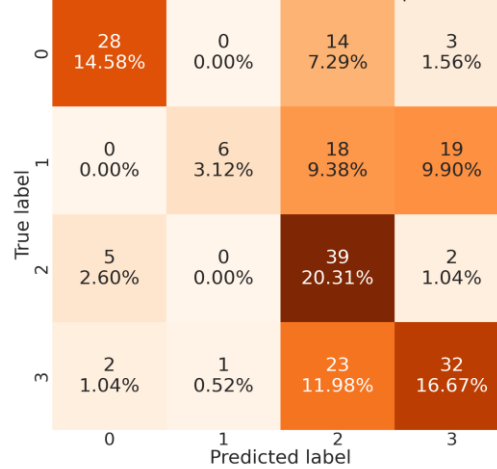
Confusion Matrix for MLP classifier with Sparse Autoencoder



Accuracy=95.000

**Figure 8.** Confusion Matrix Heatmap for MLP classifier with Sparse Autoencoder

Confusion Matrix for MLP classifier without Sparse Autoencoder



Accuracy=55.000

**Figure 9.** Confusion Matrix Heatmap for MLP classifier without Sparse Autoencoder

## 5 Conclusions and Future Scope

Sparse Autoencoder was appended with the Multi-Layer Perceptron Classifier model to produce a hybrid SAE-MLP model, which has been proved to improve speech emotion classification. On experimentation, it was observed that the Multi-Layered Perceptron did not perform well over the given speech emotion dataset and so, SAE-based automated feature engineering was employed to enhance the classification performance of the MLP. This research paper has introduced a novel hybrid model SAE-MLP for speech emotion detection, with enhanced accuracy and precision.

In the future, the speech features namely the MFCC, RAS-MFCC, LPCC, PLP, Harmonic cepstrum, would be experimented on the Speech Emotion Recognition System.

## Acknowledgment

This project is partially aided by the All India Council of Technical Education, India under the Research Progress Scheme (Ref. 8-40/RIFD/RPS/Policy-1/2017-18) - 15<sup>th</sup> March 2019.

## Declaration of Conflict of interest

The authors alone are responsible for the content and writing of the paper and they report no conflict of interest.

## References

1. Juyal S., Killa C., Singh G.P., Gupta N., Gupta V. : Emotion Recognition from Speech Using Deep Neural Network. In: Srivastava S., Khari M., Gonzalez Crespo R., Chaudhary G., Arora P. (eds) Concepts and Real-Time Applications of Deep Learning. EAI/Springer Innovations in Communication and Computing. Springer, Cham. (2021) [https://link.springer.com/chapter/10.1007/978-3-030-76167-7\\_1](https://link.springer.com/chapter/10.1007/978-3-030-76167-7_1)
2. Huang, F., Zhang, J., Zhou, C. et al. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. Landslides 17, 217–229 (2020). <https://link.springer.com/article/10.1007%2Fs10346-019-01274-9>
3. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain,: Speech Emotion Recognition Using Deep Learning Techniques: A Review, In: IEEE Access, vol. 7, pp. 117327-117345, (2019), <https://ieeexplore.ieee.org/document/8805181>
4. F. Chollet,: Building autoencoders, In: Keras (2016), <https://blog.keras.io/building-autoencoders-in-keras.html>
5. Sheena Christabel Pravin, Palanivelan, M,: A Hybrid Deep Ensemble for Speech Disfluency Classification,,In: Circuits, Systems, and Signal Processing, Springer, vol. 40, no.8, pp. 3968-3995, (July 2021). [https://www.researchgate.net/publication/349228170\\_A\\_Hybrid\\_Deep\\_Ensemble\\_for\\_Speech\\_Disfluency\\_Classification](https://www.researchgate.net/publication/349228170_A_Hybrid_Deep_Ensemble_for_Speech_Disfluency_Classification)
6. Krishnan, P.T., Joseph Raj, A.N. & Rajangam V. : Emotion classification from speech signal based on empirical mode decomposition and non-linear features. Complex Intell. Syst. 7, 1919–1934 (2021). <https://link.springer.com/article/10.1007%2Fs40747-021-00295-z>
7. Mohamad Nezami. O., Jamshid Lou. P. & Karami. M.: ShEMO: a large-scale validated database for Persian speech emotion detection. Lang Resources & Evaluation 53, 1–16 (2019) <https://link.springer.com/article/10.1007%2Fs10579-018-9427-x>
8. M. Deshpande and V. Rao,: Depression detection using emotion artificial intelligence ,In: International Conference on Intelligent Sustainable Systems (ICISS), pp. 858-862 (2017) <https://ieeexplore.ieee.org/document/8389299>
9. M. N. Stolar, M. Lech, R. S. Bolia and M. Skinner,: Real-time speech emotion recognition using RGB image classification and transfer learning, 11th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1-8, (2017) <https://ieeexplore.IEEE.org/document/8270472>
10. Sheena Christabel Pravin, Palanivelan M,: Regularized Deep LSTM Autoencoder for Phonological Deviation Assessment, In: International Journal of Pattern Recognition and Artificial Intelligence, vol. 35, no.4, p. 2152002, (March 2021).<https://www.worldscientific.com/doi/abs/10.1142/S0218001421520029>

## 10.2 Conference Certificate



### E-CERTIFICATE OF PARTICIPATION

**Vishal Balaji Sivaraman**

presented the paper titled

**Sparse Autoencoder based Speech Emotion Recognition**

authored by

**Vishal Balaji Sivaraman, Sheena Christabel Pravin, Surendaranath K, Vishal A,  
Palanivelan M., Saranya J. and Priya L**

in the International Conference on Communication and Intelligent Systems held online  
during December 18-19, 2021.

ICCIS2021/271

**Dr. Vivek Shrivastava  
(General Chair)**

**Dr. Jagdish Chand Bansal  
(General Secretary, SCRS)**

<https://iccis21.scrs.in>

**RAJALAKSHMI ENGINEERING COLLEGE**

**DEPARTMENT OF ECE**

**PROGRAM OUTCOMES (POs)**

Engineering Graduates will be able to:

**PO1 Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

**PO2 Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

**PO3 Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

**PO4 Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

**PO5 Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

**PO6 The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

**PO7 Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

**PO8 Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

**PO9 Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

**PO10 Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to

comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

**PO11 Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

**PO12 Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

### **PROGRAM SPECIFIC OUTCOMES (PSOs)**

**PSO1:**An ability to carry out research in different areas of Electronics and Communication Engineering fields resulting in journal publications and product development.

**PSO2:**To design and formulate solutions for industrial requirements using Electronics and Communication engineering

**PSO3:**To understand and develop solutions required in multidisciplinary engineering fields.

**COURSE OUTCOMES (COs)**

<b>CO1</b>	To acquire practical knowledge within the chosen area of technology for project development.
<b>CO2</b>	To identify, analyze, formulate and handle projects with a comprehensive and systematic approach.
<b>CO3</b>	To contribute as an individual or in a team in development of technical projects.

## **EC17713 – MINI PROJECT WORK**

**Project Title: Emotion Detection for Individuals with Neurological Disorders Using Speech**

**Batch Members:** Surendaranath K (180801201)

Vishal B (180801223)

Vishal Balaji Sivaraman (180801224)

**Name of the Supervisor:** Dr. Sheena Christabel Pravin

### **CO - PO – PSO matrices of course**

PO/PSO CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CO1	3	3	3	3	2	3	3	3	3	2	3	3	3	3	3
CO2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
CO3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Average	3	3	3	3	2.7	3	3	3	3	2.7	3	3	3	3	3

**Note: Enter correlation levels 1, 2 or 3 as defined below:**

**1: Slight (Low) 2: Moderate (Medium) 3: Substantial (High), If there is no correlation, put -“**

**Signature of the Supervisor**